DOCUMENT RESUME

ED 145 401 CS 003 746

AUTHOR Sandifer, Mary M.
TITLE A State Department of Education View of the Evaluation of Reading Programs.
PUB DATE Feb 76
NOTE 11p.; Paper presented at the Annual Meeting of the International Reading Association Southeastern Regional Conference (Jacksonville, Florida, February 18-21, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Data Collection; *Decision Making; Diagnostic Tests; *Educational Testing; *Evaluation Methods; Norm Referenced Tests; *Program Evaluation; *Reading Programs; Reading Tests; Testing Problems

ABSTRACT

This paper focuses on certain technical aspects of reading-program evaluation, noting the differences in decision-making requirements and data-collection procedures between local administrators and State and Federal evaluators. The paper discusses the following differences in decision-making requirements: evaluating groups or individuals, observing long-range trends or daily progress, and using diagnostic or norm-referenced tests. Procedural problems in data collection include the use of equal-interval scores, testing dates, and functional-level testing. The paper concludes that local administrators and State and Federal evaluators should be aware of each other's viewpoints toward reading-program evaluation and that they should work together to collect the proper data and to provide the best possible reading program. (RL)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED145401

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Mary M. Sandifer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper Presented at the Annual Meeting of the International Reading Association, Regional Conference
(Jacksonville, Florida, February 18-21, 1976)

A State Department of Education View of the Evaluation of Reading Programs

Mary M. Sandifer

Note: This paper focuses on certain technical aspects of project evaluation and is not intended to reflect the policies of the State Department of Education of South Carolina or of any other state.

Introduction

A very frustrating situation arises in trying to reconcile State or Federal guidelines for project evaluation and the needs and wishes of project directors and curriculum supervisors from local school districts. Often economy-minded local district personnel may wish to glean data needed for evaluation reports from existing District-wide testing programs or from the testing normally carried on in the classroom. And just as often, the reviewr of the evaluation plans must make recommendations, or insist on changes, which cause considerable inconvenience or expense to local school district staffs. This paper is prompted by personal experience in encountering the problems of local staffs trying to provide the best possible program while spending the least possible money to collect data.

The need for certain state-level evaluation requirements results from at least two causes:

1. Different types of decisions--requiring different types of data--must be made at the state level and at the local level, and

CS803746

2. To obtain reliable data, certain stringencies must be observed. The technical reasons for these stringencies are seldom understood by educators who do not have a strong background in measurement.

This paper addresses some of the most common problems arising from these differences in decision-making requirements or from the procedures required for collecting reliable, interpretable data.

Decisions Based on Evaluation Results

Do decisions pertain to groups or to individuals? This is a major and obvious consideration in planning appropriate use of evaluation data. At the classroom or school level, teachers and principals are necessarily making decisions about each student, e.g., placement and amount of progress. These decisions about individuals would be inappropriate at the state level where groups of students must be considered. The questions to be answered at the state level are of the type "Do low-income disadvantaged students advance more with Program A or Program B?" or "How do students in the high-income and low-income groups compare when taught with programmed instruction?"

Instructional decisions also differ at the classroom and at the state level. General trends and progress are major questions for state planners, while teachers and principals must decide whether to re-teach a unit or go on to the next topic; that is, whether students have mastered a particular skill. At the state level, long range trends and yearly progress, rather than weekly achievement, must be studied.

Whether you test every student, or select a sample of students to be tested, depends upon evaluation requirements at the two levels. As decisions at the state level or national level pertain to groups of students rather

than individuals, sampling of students is defensible and often may be recommended for reasons of economy or as a means of exercising more control over data collection. At the school level, student sampling is unthinkable when decisions must be made about each student.

A similar consideration pertains to content sampling. Testing every fact or every skill which students have studied during the year is seldom necessary in order to fulfill state evaluation requirements. A representative sampling of skills and facts may be tested, or terminal skills and knowledge may be tested; that is, if a student knows how to multiply he certainly has learned how to add.

The difficulty of using diagnostic tests for evaluation has been another source of frustration to local program planners who wish to use the same test for evaluation that they are using in the program. Diagnostic testing is quite necessary in carrying out a program and planning daily instruction. Evaluating a program using diagnostic tests is seldom appropriate or even possible, for they are not designed to permit measurement of progress but rather to pinpoint specific areas of difficulty.

In summary major differences in types of decisions made pertain to:

A. Groups vs. individuals

B. Long range trends vs. daily progress

C. Student sampling vs. testing a population

D. Content sampling vs. complete coverage of curriculum

E. Diagnostic tests vs. norm-referenced or mastery tests

Keeping these differences in mind should help both project directors and evaluators understand the other's viewpoint---and realize that decisions or recommendations are not arbitrary.

Procedures

Other questions which should be considered in planning the evaluation of local projects pertain to procedures needed to collect reliable data. Some current issues concern the use of equal interval scores, the date of testing, and functional level testing. Each of these is discussed below.

Equal Interval Scores.

For many years, problems resulting from the use of grade equivalent scores have occurred in evaluating programs, especially compensatory programs. The example given in Table I will illustrate this point.

Table I

Form S, Level 2

Comprehensive Test of Basic Skills/S-2 Vocabulary Subtest*

Raw Score	G.E.	PR (4.7)	PR (6.7)
0	1.0	1	1
5	1.6	2	1
10	2.7	14	5
15	3.8	33	13
20	4.7	53	23
25	5.6	71	36
30	6.8	85	51
35	8.3	96	72
40	11.9	99	98

*The CTBS/S-2 was used in this example because it is the most widely used test in South Carolina. No criticism of the test itself is suggested in this example. The CTBS has exceptionally good norms, and the same pattern would be found with almost any standardized test. Grade euivalents and percentile marks were not derived to represent equal intervals but rather to provide reference points to a normative population which is assumed to be normally distributed.

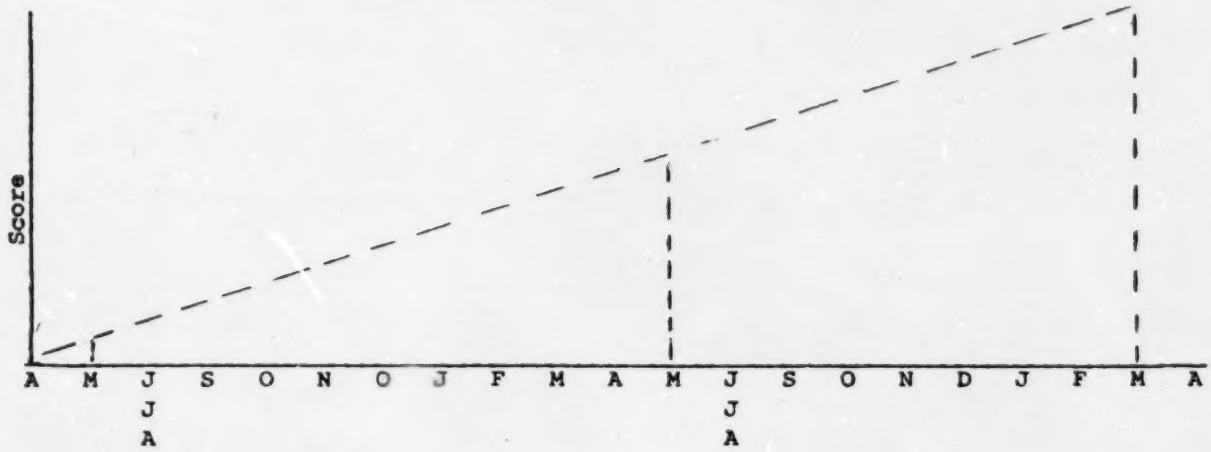
Notice that raw scores differing by five points convert into vastly different grade equivalents and percentile ranks. A change of five points at the lower range of the test, at the middle, and at the upper extremes, when converted into grade equivalents or percentile ranks will suggest vastly different amounts of student progress. Obviously, amounts of gain measured in grade equivalents or percentile ranks cannot be compared, for the amount of apparent gain depends entirely on whether the student was scoring near the floor or ceiling of the test. (Using raw scores alone, which at first glance appears to be a workable solution, is appropriate only if the items are equally difficult.) The solution to this problem is found in the use of equal interval scores called variously scale scores, standard scores, growth scores or expanded scale scores by various test publishers. These scores are derived to represent equal intervals; that is, a change of a given number of points in scale scores represents the same amount of learning regardless of where on the scale it occurs. The use of these equal interval scores for evaluation of Federal programs is now recommended by the U. S. Office of Education and may become mandatory in the near future.

Testing Date.

Evaluation data should always be gathered by administering both the pre-test and post-test as close as possible to the date on which normative data were collected. The reason for this is illustrated below. Test publishers, in establishing norms, collect data from large samples on one, or sometimes two, dates at all grade levels. For instance, the CTBS tested approximately a hundred thirty thousand (130,000) students at all grade levels in the spring. The statisticians then make straight line projections between the dates tested; (that is, from spring of the fourth grade to spring of the

fifth grade) in order to derive grade equivalents and percentile ranks for the other months which were not tested. An equal amount of progress is assumed during the nine months from September to May and one month of progress for the vacation months of June, July and August. Statisticians interpolate between these known values for testing dates to obtain the values for the months in between. Figure I below represents this.

Figure 1



X - Normative data obtained

The dotted line represents the interpolation from the dates and grades on which the data were collected. Figure 2 represents another possible learning pattern. Suppose the solid line represents the actual trend of learning in a typical student. Notice that we assume that he makes approximately equal progress during all of the months that school is in session but forgets slightly during the three summer months so that in September his level of knowledge is slightly lower than it was in May when school was dismissed.

Figure 2

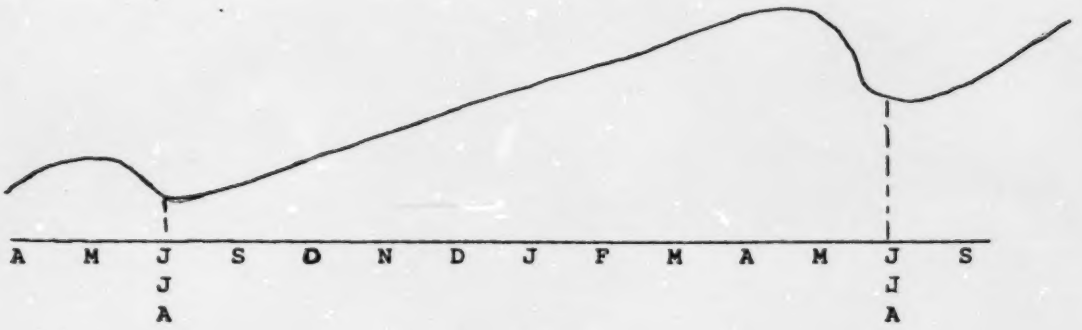
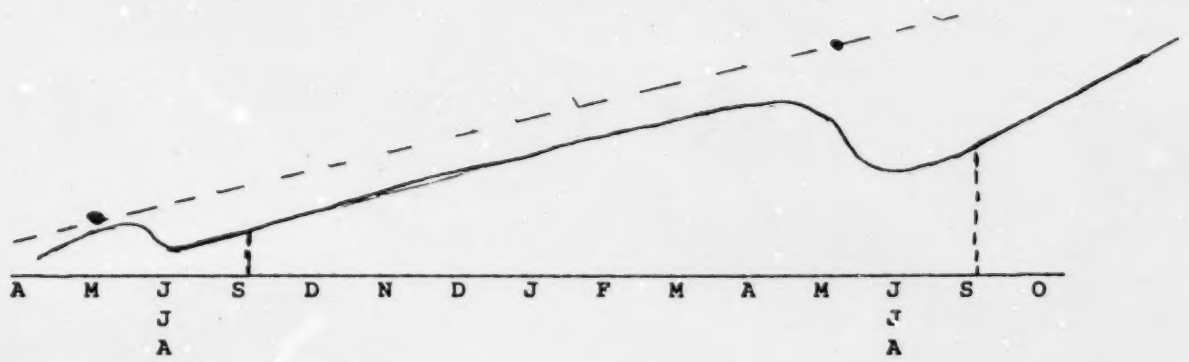


Figure 3 shows the two learning curves superimposed to illustrate what will happen if the student is tested on a date for which there are no real normative data.

Figure 3



O_1 and O_2 represent dates on which a student may be tested. Notice that if he is tested in September and his scores are interpreted according to the published norms, he will appear to be much lower than he actually is

relative to his age-mates. This will make his relative gain, when measured again in April, appear to be much larger than it actually is. In other words, when programs are evaluated by testing in fall and spring, using a test which has only spring norms, any measured gains are artificially inflated. The program will automatically appear to be much better than it actually is. Evaluators should exercise extreme caution in interpreting results of such data. The same bias will result with tests having only fall norms.

The Functional Level Testing.

Generally, in evaluating compensatory programs it is necessary to test at the child's functional level (his real reading level) rather than using a standardized test which is appropriate for his age and grade placement. The difficulty in functional level testing lies in interpreting the results relative to a national normative group, because norms generally are not available if a student is tested with a lower (or higher) level of a standardized test. Recently, with the development of scale scores and other equal interval, level-independent scores, it has become possible to interpret the results of functional level testing even though one must still exercise caution.

The argument in favor of functional level testing is simple. If a student is unable to read a test, he obviously cannot perform adequately on it. With functional level testing one will have reasonably reliable data which is difficult to interpret. Without functional level testing, one will obtain unreliable data which cannot be interpreted in any way. There is obviously little choice between the two.

In choosing a functional level test several things should be considered. Unfortunately, there are no precise guidelines or formulas to tell exactly which level will always be best for a particular situation. Factors which should be considered are:

A. The ceiling and floor of the test - The post-test score should not be too close to the floor or the ceiling but be fairly close to the middle.

B. Chance scores - The implications of chance level scores* are often overlooked by program planners and teachers utilizing test results. Often, results are reported which are at, or even below, a chance level for the particular test which is used; this is especially true of compensatory programs.

   The chance level score for any test can be found quickly by dividing the number of questions by the number of options to determine a chance level raw score. Then, a norms table will give the grade equivalent or percentile rank for that particular chance level score. If a significant number of students taking a test are functioning at or below a chance level, there should be a concerted effort to move towards functional level testing.

*Definition: Total number of questions divided by number of options for each. Assume that you are giving a one-hundred-item multiple-choice test in which each of the one hundred items has four options from which the student may choose. Now assume that you pass out answer sheets with no test booklets and tell a large class of students to guess at any answer. By chance, they would get approximately one fourth of the answers correct by guessing; that is, on each question they would have one chance out of four of guessing the right answer, even without reading the questions. This is known as a chance level score.

C. Content - This should be obvious but is often overlooked. The content of a test should be related to what the students are studying at that grade level. Also, in functional level testing it should be related to the students' interest level. In deciding whether to move to functional level testing, program planners should place heavy consideration on the actual content of the test questions at the various levels.

D. Reading Level - This is so obvious that it requires no further comment.

E. Anticipated Progress Between Pre-Test and Post-Test - Teachers may reasonably expect students to score slightly higher on a post-test than a pre-test and should consider this in deciding on the most appropriate test level to use. Unfortunately, there is no handy rule of thumb to determine how much higher a student may be expected to score on a post-test.

Concluding Comments

In planning evaluation (to quote a current popular advertisement) "There are no easy answers, only intelligent alternatives." Program planners are not likely to find an evaluation instrument which meets all of the ideal recommendations listed above. When compromises must be made, program planners should not be frustrated and should simply be aware of the potential biases and inaccuracies in the scores that they are reporting. Similarly, state and federal evaluators should assist local staff members in obtaining the best possible data and in making the most defensible interpretation of that which they do obtain.